

## **Teacher-Intervention Confound**

In some studies reviewed by the WWC, only one teacher is assigned to each condition. In particular, three different kinds of studies involve only one teacher per condition. The technical guidance discusses each case in turn. Some of these cases apply only to randomized controlled trials (RCT) while others apply to both RCTs and quasi-experimental designs (QEDs). The final case applies not only to teachers, but to any aggregated units such as classrooms, schools, or districts.

- 1) RCTs with one teacher per condition, and students randomly assigned to teachers
- 2) RCTs and QEDs with one teacher teaching both conditions, and students assigned to conditions.
- 3) RCTs with one teacher, school, or district randomly assigned to each condition and students are not randomly assigned, and similar QEDs

Finally, this guidance focuses on one specific technical issue, the confound between teacher and intervention. The study's ultimate disposition (i.e., meets evidence standards, meets evidence standards with reservations, does not meet evidence screens) also depends on how it fares on other criteria in the WWC Study Review Standards.

### **RCTs with one teacher per condition and students randomly assigned to teachers**

This part of the guidance focuses on RCTs only, and does not apply to QEDs.

1. In some studies, one teacher may teach curriculum A and a different teacher may teach curriculum B. Children are then randomly assigned to each teacher/curriculum combination.

This is indeed a randomized trial. But the estimate of the intervention's effect is problematic because the teacher and intervention are confounded. That is, the effect of teachers usually cannot be disentangled from the effect of the intervention; consequently, the estimate of the intervention's effect could be then subject to potentially serious bias.

2. The default for handling these studies is the following:

Such an RCT study should generally be downgraded or even discarded if (a) the study does not demonstrate that the confounding problem is negligible and (b) the PI and Review Team regard the potential bias in estimating effect as non-trivial.

3. In certain domains and interventions, it is possible for teacher effects to be negligible. For instance, a computer instruction program may be relatively free-standing and require little teacher engagement in the actual programmatic instruction and measurement of outcomes. In a comparison of two such computer programs, teachers might have very little effect on either condition. If the PI and Review Team agree that the study author demonstrates that teacher effects and the potential bias are negligible, then the study may

be regarded as an RCT without a teacher confound problem (that is, the study is neither downgraded nor discarded).

If the teacher has some role in implementing the intervention, but that role is limited by the nature of the intervention (e.g., predominantly computer-based), it is reasonable to assume some limited teacher effects. In this case, the study might be downgraded, but not discarded.

For interventions where the extent of teacher engagement (and therefore the possible teacher effect) is unclear, the burden of proof is on the study authors to demonstrate that teacher effects are negligible, are likely to have little or no impact on the study findings, and therefore the study should not be downgraded or discarded.

### **RCTs and QEDs with one teacher teaching both conditions and students assigned to conditions**

1. In some studies, one teacher may teach curriculum A in one class and the same teacher may teach curriculum B in a second class. Students are then randomly assigned to each class.

2. The study is a fair test of the intervention if the PI and Team believe it is reasonable to assume (a) that the teacher's ability and motivation to teach curriculum A is the same as his or her ability and motivation to teach curriculum B or (b) that effects of the teacher are negligible for this intervention (e.g., as in the example above, an intervention may require very little input on the part of a teacher). The study may provide evidence bearing on either assumption, and this should be recognized by the PI and Review Team. For instance, the study may tell the reader that the teacher is well trained in each curriculum.

This situation is analogous to some surgical trials in which the same surgeon uses two different approaches in each arm of a randomized trial. Patients are randomly assigned to each arm of the trial, but the same surgeon performs the surgery in both arms.

3. The study is not a fair test of the intervention if the PI and Review Team do not feel there is adequate basis for making any of the above assumptions. For instance, (a) the study may provide no information about the plausibility of the assumptions, and (b) the PI and Review Team regard the assumptions as implausible based on the study's contents, and (c) the PI and Review Team regard the potential bias due to teacher confound in estimating the intervention's effect as non-trivial.

4. For RCTs in which a single teacher teaches both the intervention and the control conditions, and students are randomly assigned to conditions, the WWC recommends the following default disposition.

The study should be downgraded if:

- The study author does not demonstrate equal ability and motivation of the teacher in teaching both conditions

OR

- The study author does not demonstrate negligible teacher effects for the particular intervention (if counter evidence exists)

OR

- The PI and Review Team regard the potential biases in estimating the intervention's effect as non-trivial .

The study is not downgraded if there is a strong case that teacher ability and motivation are equal in each condition or teacher effects are negligible for the particular intervention (and consequently there are no serious potential biases in the estimate of the intervention's effect). However, the PI and team should explain in the intervention report that the teacher is assumed to be equally skilled and motivated to teach in each condition.

5. QEDs are handled similarly. The study should be downgraded (i.e., discarded) if:

- The study author does not demonstrate equal ability and motivation of the teacher in teaching both conditions

OR

- The study author does not demonstrate negligible teacher effects for the particular intervention (if there is counterevidence)

OR

- The equating procedures are absent or inadequate,.

The reasons for discard should be documented and explained in the intervention report.

QEDs in which a single teacher teaches both conditions can be included in the WWC's review if the study author demonstrates that the teacher ability and motivation are equal in both conditions or teacher effects are negligible for the particular intervention. Again, this should be made explicit in the intervention report .

**RCTs with one teacher, school or district randomly assigned to each condition, and students are not randomly assigned, and similar QEDs**

1. A study may be based on two intact classrooms and their teachers, where one intact class and its teacher may be assigned randomly to condition A, and the second intact class and teacher assigned to condition B.

More generally, a study involving two aggregated intact units (e.g., classrooms, schools, or districts) may randomly assign one aggregated unit to the intervention condition and a second aggregated unit to a second condition. In the technical jargon, the aggregates are often called "clusters," "groups," or "places."

2. In each case, the unit of randomization is at the cluster (aggregate or place) level. In each case, only one unit (the teacher and her classroom, an entire school, etc.) out of two such units is randomly assigned to each treatment condition.

3. A correct statistical analysis at the level of the unit of randomization (schools or districts or classes) cannot be done without invoking untestable assumptions. This is because the number of degrees of freedom associated with statistical tests (such as t or F), confidence intervals, etc. is zero. Put another way, neither statistical significance tests nor confidence intervals can be calculated at the proper level of analysis (i.e. the level of randomization) if the study is viewed as a randomized trial. In addition, any estimate of the intervention's effect is confounded with the teacher's effect.

This design, with N=1 unit of randomization in each arm of a randomized trial, is not a good randomized design.

4. The study author may have analyzed the data at a level of units lower than the level of random assignment. For instance, the study may report an analysis based on data at the level of individual students within the randomly assigned classes, or students within the randomly assigned schools in an attempt to adjust for difference between students in different classes or schools. The study can be construed as a QED if the analysis was done this way.

5. The default disposition for such RCTs is as follows:

Such a study should generally be downgraded by the PI and Team. Depending on the study design, analysis, and the assumptions, the study may have been analyzed as a QED. If it does not meet the standards for a QED, it should be discarded. The PI and Review Team should document the reason for the discard.

6. QEDs in which schools (or other entities, such as classes or districts) are confounded with interventions are problematic in that the effects of schools and effects of the intervention usually cannot be disentangled, and the assumption that the school effects are equal is usually not plausible. Further, post facto matching of students or (equivalently) statistically equating is often suspect. For instance, if the schools differ appreciably in their location and characteristics of students (New York City versus rural Iowa), no amount of matching or statistical equating is likely to assure that the groups of children that are finally selected as being comparable within schools will indeed yield a statistically unbiased estimate of the effect of the intervention.

The WWC regards the assumption as patently implausible in the study context (or regard the equating as patently inadequate) and should then downgrade the study and discard it. Reasons for the discard must be given. However, the PI and Review Team may include a study of this type if they can provide compelling evidence that the required assumptions have been met.

*Bibliography*

Wolins, Leroy. (1982) *Research Mistakes in the Behavioral and Social Sciences*. Ames, Iowa: Iowa State University Press.